# Disentangling neural representations of value and salience in the human brain

Kahnt, Thorsten ; Park, Soyoung Q ; Haynes, John-Dylan ; Tobler, Philippe N

Abstract: A large body of evidence has implicated the posterior parietal and orbitofrontal cortex in the processing of value. However, value correlates perfectly with salience when appetitive stimuli are investigated in isolation. Accordingly, considerable uncertainty has remained about the precise nature of the previously identified signals. In particular, recent evidence suggests that neurons in the primate parietal cortex signal salience instead of value. To investigate neural signatures of value and salience, here we apply multivariate (pattern-based) analyses to human functional MRI data acquired during a noninstrumental outcome-prediction task involving appetitive and aversive outcomes. Reaction time data indicated additive and independent effects of value and salience. Critically, we show that multivoxel ensemble activity in the posterior parietal cortex encodes predicted value and salience in superior and inferior compartments, respectively. These findings reinforce the earlier reports of parietal value signals and reconcile them with the recent salience report. Moreover, we find that multivoxel patterns in the orbitofrontal cortex correlate with value. Importantly, the patterns coding for the predicted value of appetitive and aversive outcomes are similar, indicating a common neural scale for appetite and aversive values in the orbitofrontal cortex. Thus orbitofrontal activity patterns satisfy a basic requirement for a neural value signal.

DOI: https://doi.org/10.1073/pnas.1320189111

**BIOLOGICAL SCIENCES – Neuroscience**
**SOCIAL SCIENCES – Psychological and Cognitive Sciences**

# Disentangling neural representations of value and salience in the human brain

Thorsten Kahnt[1,2], Soyoung Q Park[1,3,2], John-Dylan Haynes[2,4], Philippe N. Tobler[1]

*[1]Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich, 8006 Zürich, Switzerland*
*[2]Bernstein Center for Computational Neuroscience, Charité – Universitätsmedizin Berlin, 10115 Berlin, Germany*
*[3]Department of Psychology, University of Lübeck, 23562 Lübeck, Germany*
*[4]Berlin Center for Advanced Neuroimaging, Charité – Universitätsmedizin Berlin, 10115 Berlin, Germany*

**Corresponding author**
Thorsten Kahnt
Laboratory for Social and Neural Systems Research
Department of Economics
University of Zurich
Blümlisalpstrasse 10
8006 Zürich, Switzerland
Phone: +41 44 634 37 40
Email: thorsten.kahnt@econ.uzh.ch

## Keywords
Reward; punishment; decision-making; attention; MVPA

Value and salience in the human brain

**Abstract**

A large body of evidence has implicated the posterior parietal and orbitofrontal cortex in the processing of value. However, value perfectly correlates with salience when appetitive stimuli are investigated in isolation. Accordingly, considerable uncertainty has remained about the precise nature of the previously identified signals. In particular, recent evidence suggests that neurons in the primate parietal cortex signal salience instead of value. To investigate neural signatures of value and salience, here we apply multivariate (pattern-based) analyses to human fMRI data acquired during a non-instrumental outcome-prediction task involving appetitive and aversive outcomes. Reaction time data indicated additive and independent effects of value and salience. Critically, we show that multivoxel ensemble activity in the posterior parietal cortex encodes predicted value and salience in superior and inferior compartments, respectively. These findings reinforce the earlier reports of parietal value signals and reconcile them with the recent salience report. Moreover, we find that multivoxel patterns in the orbitofrontal cortex correlate with value. Importantly, the patterns coding for the predicted value of appetitive and aversive outcomes are similar, indicating a common neural scale for appetite and aversive values in the orbitofrontal cortex. Thereby, orbitofrontal activity patterns satisfy a basic requirement for a neural value signal.

**Significance Statement**

The value and salience of predictive cues are important signals for regulating approach-avoidance behavior and attentional processing, respectively. However, both signals are often confounded in studies of decision-making. Indeed, recent results suggest that neural signals in the primate posterior parietal cortex (PPC) which were previously thought to encode value are actually reflecting salience. This finding has created considerable uncertainty about previously identified value signals. Here we experimentally dissociate value and salience, and use pattern-based fMRI to demonstrate distinct encoding of both signals in the PPC, thereby reinforcing the earlier reports of value in the PPC. Moreover, we show that the orbitofrontal cortex encodes the predicted value of appetitive and aversive outcomes on a common neural scale.

## Introduction

The value of predictive cues can be used to guide approach-avoidance behavior. Approach and avoidance responses are proportional to the appetitive (positive) and aversive (negative) value of the cues, respectively. On the other hand, based on empirical and theoretical considerations (1-3) the absolute value (i.e. the salience) of a cue determines the amount of attention that a stimulus captures to facilitate further processing. Hence, in contrast to value, salience increases not only with the magnitude of reward but also with the magnitude of punishment (4, 5).

Electrophysiological recordings in animals suggest that value is encoded in the firing rates of posterior parietal and orbitofrontal neurons (6-18). There is also a large body of evidence from human imaging studies suggesting signatures of appetitive value in these regions (19-36). However, value and salience are perfectly correlated when appetitive stimuli are investigated in isolation (37). That is, if a signal increases with increasing reward, we need to know how it behaves with increasing punishments in order to decide whether it is coding for value or salience. Specifically, if the signal decreases with increasing punishment, it truly reflects value. However, if the signal also increases with increasing punishment, it reflects salience (see **Fig. 1*C***). Thus, value signals identified using only appetitive (or only aversive) stimuli could be explained equally well in terms of value or salience.

Indeed, neurons in the lateral intraparietal area (LIP), which have long been thought to signal decision values (6-9), have recently been shown to signal salience (37). This result has created considerable uncertainty regarding previous findings on the neural coding of value, not only in the posterior parietal cortex (PPC) but also in the orbitofrontal cortex (OFC). To assess the nature of anticipatory value and salience signals in the human PPC and OFC, here we use a non-instrumental outcome-prediction task and multivoxel pattern-based fMRI. This analysis technique combines the activity of multiple voxels, and can be used to reveal signals encoded in intercalated neuronal populations (see **SI**

**Discussion**). In order to dissociate value and salience signals, the current task involves distinct stimuli predicting small or large appetitive or aversive outcomes.

We carry out three complementary analyses on the fMRI data aiming at where and how value and salience signals are represented. First, we identify brain regions carrying information about the two signals. This is done by searching for multivoxel response patterns that code for one or the other variable. Second, in regions encoding value or salience, we test whether these multivoxel patterns code for value differences within the appetitive and the aversive domain (thus reflecting graded value and not only categorical valence), and whether appetitive and aversive cues contribute similarly to the observed neural encoding. Third, we ask whether the multivoxel ensemble codes of value or salience are similar for appetitive and aversive values. In other words, we test whether we can predict the value of an aversive cue from multivoxel response patterns coding for the value of appetitive cues.

## Results

*Behavioral results.* We employed a simple non-instrumental outcome-prediction task (**Fig. 1*A***) in which visual cues deterministically (100% cue-outcome contingency) predict the gain or loss of small or large amounts of money (i.e. 0.50 € or 5.00 €). Here we assume value to linearly increase with nominal monetary magnitude, which usually holds for small amounts or intervals (38) (see **SI Discussion**)**.** Two sets of cues were associated with the four possible outcomes (−5.00 €, −0.50 €, 0.50 € and 5.00 €; **Fig. 1*B***), such that each outcome was predicted by two different visual cues. We used two sets in order to control for the sensory properties of the cues (29). Subjects (n = 30) had to indicate the predicted outcome after a variable delay and before the actual outcome was shown. On average, subjects were correct in predicting the outcome on almost every trial (average %-correct = 95.54; one sample t-test against chance [4 options, chance = 25%], $t_{29}$ = 132.60, $p < 0.001$). However, note that the cue-outcome pairing was purely non-instrumental and thus outcomes were independent of the correctness of the response.

By using both appetitive and aversive cues, this task allows for linearly independent (i.e. uncorrelated) levels of value and salience (**Fig. 1*C***). We used response time (RT) to search for behavioral effects of value and salience. Specifically, in order to estimate and compare the behavioral effects of value and salience, we applied subject-wise multiple regression models (see **Materials and Methods**). On average, we observed significant negative regression coefficients for both value (one sample t-test, $t_{29} = -2.82$, $p = 0.009$) and salience ($t_{29} = -2.72$, $p = 0.011$), indicating that high levels of value and salience led subjects to respond faster. Moreover, the effects of value and salience on RT did not differ (paired t-test on standardized regression coefficients, $t_{29} = 0.29$, $p = 0.78$), suggesting that both variables affected behavior to a comparable degree. The independent influence of value and salience on RT should result in a magnitude by valence interaction and a main effect of magnitude, and our data confirmed these predictions (see **Fig. S1** and **SI Results**).

*Identifying brain regions coding for value and salience.* First, we searched for brain regions coding for the predicted value of the cues independent of their sensory properties. For this we employed a searchlight cross-decoding approach using linear support vector regression (SVR) and leave-one out cross-validation (see **Materials and Methods**). In brief, for each searchlight (radius 3 voxels), we trained a SVR based on the multivoxel response patterns evoked by the cues in set I (using the value of appetitive and aversive cues as labels, i.e. −5.00 €, −0.50 €, 0.50 € and 5.00 €) and predicted the value of the cues in set II based on their corresponding response patterns (**Fig. 2*A***). We also performed the same analysis in the opposite direction by training on cues from set II and testing on cues from set I (results represent the average). Across subjects, we find significant information about the predicted value of the cues in the central OFC (**Fig. 2*B*,** Montreal Neurological Institute (MNI) coordinates [x, y, z], [−21, 53, −14], $t_{29} = 3.83$, $p_{FWE\text{-}corr} = 0.048$). Moreover, we also find significant information about value in superior regions of the PPC along the intraparietal sulcus (IPS) (**Fig. 2*C*,** left IPS, [−48, −61, 46], $t_{29} = 6.70$, $p_{FWE\text{-}corr} < 0.001$; right IPS, [39, −73,

43], $t_{29}$ = 4.65, $p_{FWE\text{-}corr}$ = 0.015). Thus, multivoxel patters in these regions can be used to make predictions about the value of the predictive cues (**Fig. 2*D***).
Second, we searched for brain regions which encode the salience of the cues independent of their sensory properties. For this we applied the pattern-based analysis as described above, but this time we used salience (absolute value) as label for training and testing the SVR (i.e. 5.00 €, 0.50 €, 0.50 € and 5.00 €) (**Fig. 3*A***). We found significant information about the salience of the cues in the PPC, specifically in inferior regions such as the temporoparietal junction (TPJ) (**Fig. 3*B*,** left TPJ, [−60, −43, 31] $t_{29}$ = 4.65, $p_{FWE\text{-}corr}$ = 0.019; right TPJ [60, −49, 34], $t_{29}$ = 5.58, $p_{FWE\text{-}corr}$ = 0.002) but also a trend in the right IPS ([24, −52, 58], $t_{29}$ = 4.19, $p_{FWE\text{-}corr}$ = 0.056) and more medial areas extending into the precuneus, [−12, −73, 43] $t_{29}$ = 4.36, $p_{FWE\text{-}corr}$ = 0.037). Finally, we also find significant salience information in the anterior cingulate cortex (ACC) (**Fig. 3*C***, [6, 41, 25], $t_{29}$ = 4.10, $p_{FWE\text{-}corr}$ = 0.029). Taken together, the results of these analyses show that value and salience signals are both encoded in the PPC, albeit in different sub-regions spanning a superior-inferior gradient (**Fig. 4** and **Fig. S2**). The superior PPC encodes primarily value but also shows a trend for salience, whereas the inferior PPC encodes salience only.

*Decoding the value and salience of appetitive and aversive cues.* In a post-hoc analysis, we addressed three further issues. First, we asked whether the neural patterns identified above indeed code for the graded value of both appetitive and aversive predicted outcomes, and not just the general difference between aversive and appetitive domains (categorical valence). That is, we tested whether these patterns encode information not only about the sign of the predicted outcome (i.e. appetitive vs. aversive) but also about the degree to which predicted outcomes are appetitive or aversive (i.e. low vs. high), as would be expected from a value coding region. Second, we wanted to rule out that only appetitive or only aversive predicted outcomes contributed to the encoding of either value or salience. Third, we asked whether appetitive and aversive values contributed differentially to the neural encoding of value.

To address these issues, we used the SVR model which was trained on the value of all cues from set I, and tested it separately on the value of appetitive and aversive cues from set II (see **Fig. 5*A***), and vice versa by training on set II, and separately testing on appetitive and aversive cues from set I (results are averaged across both directions). This procedure results in two accuracy measures, one reflecting the encoding of appetitive values, and the other the encoding of aversive values (**Fig. 5*A***). We performed this analysis on the individual value-coding patterns identified above using the individual peak-accuracy searchlights (radius 3 voxels) in value-coding regions. In particular, we used the individual peak-accuracy searchlights in a 12 mm sphere surrounding the group peak voxel in OFC, left and right IPS (changing the size of this search sphere to 8 - 14 mm led to qualitatively similar results). All accuracies for testing the SVR model individually on appetitive and aversive cues were significant in all regions (one sample t-tests, all ps < 0.001). This indicates that the multivoxel response patterns identified above encode not only the sign of the predicted outcome, but also the degree to which predicted outcomes are appetitive or aversive. Also, these results demonstrate that the value codes are indeed not driven only by aversive cues or only by appetitive cues.

Next, we tested whether appetitive and aversive values contributed differentially to the neural encoding of value. A two-way (region-by-valence) ANOVA with repeated measures on accuracy revealed no significant main effect of region ($F_{2,58}$ = 1.79, p = 0.18), valence ($F_{1,29}$ = 2.14, p = 0.15), and no significant region-by-valence interaction ($F_{2,58}$ = 0.54, p = 0.59). This finding suggests that appetitive and aversive values are not differentially encoded. However, it is based on a null result, and thus should not form the basis of strong conclusions.

We performed a parallel analysis in the salience-coding regions defined above (ACC, left and right TPJ) to estimate the encoding of appetitive and aversive salience information (**Fig. 5*B***). In all regions, accuracies for appetitive and aversive predicted outcomes were significant (all ps < 0.001), demonstrating that salience encoding was not driven by appetitive or aversive cues only.

Moreover, a two-way (region-by-valence) ANOVA with repeated measures on accuracy revealed no significant effect of region ($F_{2,58}$ = 0.84, p = 0.45), valence ($F_{1,29}$ = 1.84, p = 0.19) or their interaction ($F_{2,58}$ = 0.97, p = 0.39), suggesting that salience is not differentially encoded in appetitive and aversive domains. Note though that this finding is also based on a null result and should be treated with caution.

*Common neural scale for appetitive and aversive value.* In a further post-hoc analysis we tested whether the value of appetitive and aversive cues is represented by a similar neural code. If the brain represents values spanning the full range from negative (aversive) to positive (appetitive) levels on one common neural scale, it should be possible to decode the value of aversive cues based on the knowledge we have about the neural encoding of appetitive cues. In other words, the difference between something *slightly* good (0.50 €) and something *very* good (5.00 €) should result in a similar multivoxel response pattern as the difference between something *very* bad (−5.00 €) and something *slightly* bad (−0.50 €).

We tested this idea in the individual value-coding multivoxel patterns as defined above, using the same individual peak-accuracy searchlights (radius 3 voxels) in a 12 mm sphere surrounding the group peak voxel in OFC, left and right IPS (again, changing the size of this search sphere to 8 - 14 mm led to qualitatively similar results). In particular, we trained a SVR on multivoxel activity patterns from cues with positive values (5.00 € > 0.50 €) and tested it on activity patterns from cues with negative value (−0.50 € > −5.00 €) (see **Fig. 6*A***), and vice versa (results represent the average). This cross-valence value decoding was significant in the OFC (one sample t-test, $t_{29}$ = 2.45, p = 0.02), indicating that differences in values above and below zero entail similar multivoxel ensemble codes. In other words, neural value representations in the OFC are invariant to the valence of the expected outcome. By contrast, this invariance was observed in the bilateral IPS (left, $t_{29}$ = −1.23, p = 0.23; right, $t_{29}$ = 1.31, p = 0.20; **Fig. 6*A***) which is likely to be caused by the simultaneous presence of value and salience

signals in this region (see **Fig. 4** and **Fig. S2**, and **SI Results** for analyses ruling out alternative explanations).

We also performed the corresponding analysis for salience in salience-coding regions (ACC, left and right TPJ) in order to test whether salience encoding for appetitive and aversive cues is similar. We trained a SVR on activity patterns from appetitive cues (5.00 € > 0.50 €) and tested it on activity patterns from aversive cues (−5.00 € > −0.50 €), and vice versa. We find significant cross-valence salience encoding in all three brain regions (one sample t-tests, all ps < 0.001, **Fig. 6*B***) and no main effect of region (one-way ANOVA; $F_{2,58} = 0.24$, $p = 0.79$). This shows that in all salience-coding regions, high vs. low appetitive and aversive values are represented by similar activity patterns.

## Discussion

*Representations of value and salience in the PPC.* In the current experiment, we have shown that multivoxel activity patterns in the PPC correlate with both value and salience. Value signals in the PPC have long been investigated in primates (6-9). For instance, in monkeys engaged in saccadic decisions, the activity of LIP neurons scales with the expected reward associated with saccadic targets in the neurons' response fields (6). Moreover, even if the cue in the response field does not provide action information, these neurons change their activity according to whether or not the cue predicts reward (39). Our results further reinforce the notion of value coding in the PPC by revealing such signals even in a non-instrumental task.

However, the view that LIP neurons encode value has recently been challenged by a study which used both appetitive and aversive decision outcomes (37). The authors showed that single LIP neurons fire strongly to both highly appetitive and highly aversive outcomes. This firing profile is incompatible with a value account, but suggests that LIP neurons are actually coding for the salience of the option in their response fields. Thus, instead of containing information about the value and the particular type of response (approach vs. avoidance), these results suggest that LIP is signaling the importance of a cue

independent of its valence. By using an unbiased whole-brain approach, we reconcile these seemingly contradictory findings, and demonstrate that both value and salience signals are present in the PPC. We show that in line with previous results, the inferior PPC including the TPJ encodes salience (40), whereas the superior PPC encodes primarily value but also shows a trend for salience. Thus, our results suggest that the PPC encodes the importance of cues (37), and is involved not only in shifting attention and accumulating further information (41), but also in guiding utility-maximizing behavior (6). Thus, there are indeed value signals in the PPC, and the conclusions of the previous studies reporting value-coding neurons (6-9) were essentially correct. However, because the designs of these studies did not include aversive stimuli, they may have misclassified salience-coding neurons as value-coding neurons and therefore overestimated the prevalence of value-coding neurons.

Furthermore, our results suggest that instead of restricting oneself to the LIP, a more comprehensive coverage of additional PPC regions with neurophysiological methods assessing both appetitive and aversive stimuli may be warranted. However, an important issue to consider when comparing neuroimaging with neurophysiology is the lack of correspondence between BOLD signals and neural spiking. BOLD signals more closely follow input into, and local processing within a region, rather than the spiking output (42, 43). Thus, because salience signals can be constructed from value input (but not vice versa), it is possible that the value information we observe in the PPC may reflect input or local processing that is not represented in the spiking output of this region. Moreover, compared to neurophysiology, the spatial resolution of neuroimaging is limited and it is conceivable that intercalated populations of value and salience coding neurons could be detected with electrodes but not with scanners. However, even though the link between multivoxel patterns and neurophysiology is not fully understood (see **SI Discussion**), using pattern-based analyses revealed results not obtained using univariate analyses (see **SI Results**).

Finally, value signals in LIP have been questioned based on the finding of phasic, cue-locked signals being related to salience rather than value (37). However, this report was controversial because delay-period activity, which is a classic marker of value- or intention-related activity in LIP (6-9), was not seen in their data (44). It is possible that the authors recorded from a different region than previous studies, and that value and salience signals co-exist in the PPC as suggested by our current findings and a recent inactivation study (45).

*Representations of value in the OFC.* We find value to be represented in the central OFC (see **SI Discussion** on localization). This finding is in line with a large number of animal recording (10-18) and human imaging studies (20-36). By using a non-instrumental task with both appetitive and aversive outcomes our results extend and inform these findings in several ways. First we show that OFC represents value even in the absence of decisions, that is, independent of action values, chosen values and other choice signals. Second, by using both appetitive and aversive predicted outcomes we demonstrate that these anticipatory signals indeed code for value rather than salience.

Moreover, using pattern-based analysis allowed us to show that the positive value of appetitive cues is represented on the same neural scale as the negative value of aversive cues. Specifically, the difference between two expected outcomes, for which one is better (i.e. more desirable) than the other, is represented by the same multivoxel pattern independent of whether the two outcomes are appetitive or aversive. These results were achieved by training and testing within-subject multivariate models on appetitive and aversive values, respectively. By doing so, we explicitly tested for similar neural codes of appetitive and aversive values and demonstrated that appetitive and aversive values are indeed encoded by the same multivoxel response pattern. Importantly, this finding suggests the presence of neurons that encode appetitive and aversive outcomes on a common value scale (i.e. neurons that consistently change their activity with increasing reward *and* decreasing punishment). Indeed, individual neurons showing such consistent activity changes with the value of

both rewarding and punishing outcomes have been identified in a very similar region of the monkey central OFC (14).

*Definitions of salience.* In our experiment, cues deterministically predicted the appetitive and aversive outcomes of different magnitude, and salience was defined as the absolute (unsigned) value (i.e. magnitude) of the outcomes (4, 5, 40). Note though that salience may be defined not only through magnitude but also through probability (1-3). In these views, particularly reliable ($p = 0$, $p = 1$) predictors of reward have different levels of salience than unreliable ($p = 0.5$) predictors (2). Regardless of the particular definition, the salience of a cue determines the amount of attention that is recruited for further processing and learning (2, 46). Unfortunately, direct comparisons between probability- and magnitude-based salience concepts are lacking, but as in previous experiments (40, 47), here we show that magnitude-related increases in attention are reflected in faster responding for more salient cues.

Despite the similar effects on behavior, in contrast to our findings, previous research showed probability-based uncertainty (48) and probability-based salience signals in the OFC (47). Probability based salience was encoded by overlapping neuronal populations with opposing coding schemes, and while they might have been missed in standard univariate BOLD analyses, our pattern-based approach should in principle be sensitive to such responses (49, 50) (**SI Discussion**). Even though we cannot draw firm conclusions from these negative results, we believe that this discrepancy is likely to result from differences in how salience was defined (probability vs. magnitude), rather than from methodological differences (single-cell vs. BOLD responses).

*Conclusion*. In summary, here we used cues predictive of appetitive and aversive outcomes and showed that the PPC encoded their value and salience in superior and inferior compartments, respectively. The co-occurrence of value and salience signals in the PPC mitigates discrepancies between previous single-cell recording experiments. Moreover, we have shown that the OFC encodes

appetitive and aversive values, and represents both values on a common neural scale. Such a common scale is of fundamental importance for economic decision making as it enables computations across the entire range of possible values.

**Materials and Methods**

*Subjects.* Thirty right-handed subjects (15 male; 24 ± 0.59 years old, mean ± SEM) with normal or corrected-to-normal vision participated in the experiment. The study was approved by the local ethics review board of the Humboldt-Universität zu Berlin, and subjects provided informed consent to participate.

*Stimuli and task.* To study neural representations of value and salience, we used a non-instrumental outcome-prediction task (**Fig. 1*A***) where appetitive and aversive outcomes (gains and losses of 0.50 € and 5.00 €) were deterministically (i.e. with 100% cue-outcome contingency) predicted by two sets of visual cues, resulting in a total of 8 cues (**Fig. 1*B***). Associations between cues and outcomes were randomized across subjects. In each trial of the task, one of the 8 cues was shown for 2 seconds. After a variable interval (4-8 s) subjects had to indicate the outcome that is predicted by the cue. This was done by pressing one of 4 buttons (left middle, left index, right index or right middle finger) corresponding to the position of the correct outcome on a response mapping screen (RMS). Importantly, to prevent preparatory motor signals during the cue interval, the positions of the 4 outcomes on the RMS were randomized in each trial. After the rating, the outcome was shown to the subject for 2 seconds. In each of the 6 scanning runs, each of the eight cue-outcome pairings was shown 5 times resulting in 40 trials per run. Subjects received 25 € for participation and were informed that in the end of the experiment they will randomly pick one trial (by throwing a die to select the run, and by drawing the trial number from an urn) and the outcome of the selected trial will be added (appetitive outcomes) or subtracted (aversive outcomes) from their total payment. Thus, each trial outcome had the same probability of getting realized.

Before scanning, subjects performed several training sessions. First, they were familiarized with the RMS. Second, subjects performed a classical conditioning session to learn the associations between the eight cues and the four outcomes. Finally, they performed a practice version of the actual scanner task outside the magnet. During conditioning, practice and scanner sessions all cues (and thus outcomes) were pseudo-randomly intermixed. Cues were separated into sets only for analysis purposes (see below). In the last practice session, before subjects went into the scanner, average performance was at 88.83% ($t_{29}$ = 51.55, $p < 0.001$, one sample t-test against chance = 25%). Also the RT data showed the same pattern as in the scanner. In particular, there were significant and negative effects of value ($t_{29}$ = -3.56, p = 0.0013) and salience ($t_{29}$=-3.37, p = 0.0022), which did not differ significantly ($t_{29}$ =0.55, p = 0.59). This demonstrates that subjects had already learned the cue-outcome associations before entering the scanner.

*Behavioral data analysis.* We estimated the effects of value and salience on RT using single-subject multiple regression models (21). Specifically, we simultaneously regressed trial-by-trial RT against z-standardized regressors of value and salience on the single-subject level. The resulting standardized regression coefficients reflect the independent effects of value and salience on RT (i.e. how much variance in the RT data is explained by value and salience, respectively). Please note that given their orthogonality, value and salience can independently affect RT. The regression coefficients were then tested individually for significance on the group-level by using one-sample t-tests. To compare the regression coefficients corresponding to the effects of value and salience on RT, we used a paired-sample t-test.

*fMRI data acquisition and preprocessing.* Functional MRI data was acquired on a 3-Tesla Siemens Trio (Erlangen, Germany) scanner equipped with a 12-channel head coil. In each of the 6 scanning runs 310 volumes were acquired (TR = 2 s, TE = 25 ms, 35 slices, ascending order, 3 mm thick, 0.75 mm gap, FOV 192 x

192 mm, matrix 64 x 64 yielding an in plane resolution of 3 x 3 mm). Preprocessing was performed by using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK) and included slice-time correction, realignment, and spatial normalization to a standard MNI template (resampling to 3 mm isotropic voxels). Unsmoothed images were used for the multivoxel pattern analysis, whereas spatial smoothing with a Gaussian kernel of 8 mm full width at half maximum (FWHM) was applied prior to the univariate analysis

*Multivoxel pattern analysis.* We used a searchlight decoding approach that allows whole-brain information mapping without potentially biasing voxel selection (51, 52) in combination with linear kernel SVR (53, 54). In a first step, for each subject and each run, a general linear model (GLM) was applied to the preprocessed functional imaging data. The GLM contained 8 regressors for the onsets of the 8 different cues (two sets of cues predicting −5.00 €, −0.50 €, 0.50 € or 5.00 €) and 4 regressors for the onsets of the 4 different outcomes (−5.00 €, −0.50 €, 0.50 € or 5.00 €), respectively (all convolved with a canonical hemodynamic response function), as well as six regressors accounting for variance induced by head motion. The voxel-wise parameter estimates of the first 8 regressors represent the response amplitudes to each of the 8 cues in each of the 6 scanning runs.

In a second step, these parameter estimates were used as input for two SVR decoding analyses involving either the value or the salience of the cues as labels. The SVR was performed by using the LIBSVM implementation (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) with a linear kernel and a preselected cost parameter of c = 0.01. It is important to note that we used a *linear* kernel which is not at risk of misclassifying neural value representations as salience by exploiting the v-shaped relationship between value and salience like *nonlinear* SVR (e.g. radial basis function) would do. In other words, the linear SVR used here which is trained to identify salience signals will perform at chance when only value information is present in the data. For each searchlight (all voxels within a radius of 3 voxels surrounding the central voxel), we performed a 6-fold leave-

one out cross-validation procedure. In each fold, training was based on data from cues from set I in 5 scanning runs (e.g. runs 1-5) and prediction accuracy was obtained in the independent 6th scanning run (run 6) based on cues from set II. This procedure was repeated 6 times, with each time leaving out a different scanning run for training the SVR, and testing it on this left-out scanning run. By using different cue sets to train and test the SVR, we ensured that information about value is not confounded by information about the visual features of the cues (29). The prediction accuracy assigned to the central searchlight voxel was defined as the average Fisher's z-transformed correlation coefficient between the actual labels of the independent test data set and the labels predicted by the SVR model. Because correlation coefficients are computed based on model predictions in the independent test data, and not on model fits in the training data, this cross-validation procedure is completely insensitive to potential noise fitting (i.e., overfitting) in the training data (55). For each subject, depending on whether value or salience is used as label, this method results in three-dimensional maps of locally distributed information about value or salience.

To identify brain regions where individual searchlights, containing information about value and salience overlapped significantly, we performed group-level analyses (n = 30 subjects) by using voxel-wise one sample t-tests on smoothed accuracy maps (6 mm FWHM). We applied a statistical threshold of $p < 0.05$, corrected for multiple comparisons (familywise error rate, $p_{FWE\text{-}corr} < 0.05$). Based on a priori hypotheses regarding encoding of value and salience, correction was performed within the following anatomical regions of interest from the AAL (automated anatomical labeling) atlas: OFC (superior orbital gyrus, middle orbital gyrus and inferior orbital gyrus), PPC (superior parietal gyrus, inferior parietal gyrus, supramarginal gyrus and angular gyrus) and ACC (anterior cingulum). For display purposes, all corrected results are presented at $p_{FWE\text{-}corr} < 0.05$ and $p_{uncorr} < 0.001$.

**Acknowledgements**

We thank F. Imamoglu for assistance in collecting data, M. Murusidze for recruiting subjects, and S. Hetzer for help with the scanning sequence. This work was supported by the Swiss National Science Foundation (Grants PP00P1_128574 and CRSII3_141965), the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (Grant 01GQ0411), and the Swiss National Centre of Competence in Research in Affective Sciences. We acknowledge also the Neuroscience Center Zurich.

## References

1. Mackintosh NJ (1975) Theory of Attention - Variations in Associability of Stimuli with Reinforcement. *Psychol Rev* 82(4):276-298.

2. Esber GR, Haselgrove M (2011) Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proc Biol Sci* 278:2553-2561.

3. Pearce JM, Hall G (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev* 87:532-552.

4. Bromberg-Martin ES, Matsumoto M, Hikosaka O (2010) Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron* 68:815-834.

5. Matsumoto M, Hikosaka O (2009) Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* 459:837-841.

6. Platt ML, Glimcher PW (1999) Neural correlates of decision variables in parietal cortex. *Nature* 400:233-238.

7. Yang T, Shadlen MN (2007) Probabilistic reasoning by neurons. *Nature* 447:1075-1080.

8. Sugrue LP, Corrado GS, Newsome WT (2004) Matching behavior and the representation of value in the parietal cortex. *Science* 304:1782-1787.

9. Louie K, Glimcher PW (2010) Separating value from choice: delay discounting activity in the lateral intraparietal area. *J Neurosci* 30:5498-5507.

10. Schoenbaum G, Chiba AA, Gallagher M (1998) Orbitofrontal cortex and basolateral amygdala encode expected outcomes during learning. *Nat Neurosci* 1:155-159.

11. Tremblay L, Schultz W (1999) Relative reward preference in primate orbitofrontal cortex. *Nature* 398:704-708.

12. Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. *Nature* 441:223-226.

13. Padoa-Schioppa C, Assad JA (2008) The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. *Nat Neurosci* 11:95-102.

14. Morrison SE, Salzman CD (2009) The convergence of information about rewarding and aversive stimuli in single neurons. *J Neurosci* 29:11471-11483.

15. Kobayashi S, Pinto de CO, Schultz W (2010) Adaptation of reward sensitivity in orbitofrontal neurons. *J Neurosci* 30:534-544.

16. Kennerley SW, Behrens TE, Wallis JD (2011) Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nat Neurosci* 14:1581-1589.

17. Wallis JD (2012) Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nat Neurosci* 15:13-19.

18. Luk CH, Wallis JD (2013) Choice Coding in Frontal Cortex during Stimulus-Guided or Action-Guided Decision-Making. *J Neurosci* 33:1864-1871.

19. Hare TA et al. (2011) Transformation of stimulus value signals into motor commands during simple choice. *Proc Natl Acad Sci U S A* 108:18120-18125.

20. FitzGerald TH, Friston KJ, Dolan RJ (2012) Action-specific value signals in reward-related regions of the human brain. *J Neurosci* 32:16417-23a.

21. Hunt LT et al. (2012) Mechanisms underlying cortical activity during value-guided choice. *Nat Neurosci* 15:470-473.

22. O'Doherty JP et al. (2001) Abstract reward and punishment representations in the human orbitofrontal cortex. *Nat Neurosci* 4:95-102.

23. Gottfried JA, O'Doherty J, Dolan RJ (2003) Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* 301:1104-1107.

24. Kim H, Shimojo S, O'Doherty JP (2006) Is Avoiding an Aversive Outcome Rewarding? Neural Substrates of Avoidance Learning in the Human Brain. *PLoS Biol* 4:e233.

25. Plassmann H, O'Doherty J, Rangel A (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J Neurosci* 27:9984-9988.

26. Tom SM, Fox CR, Trepel C, Poldrack RA (2007) The neural basis of loss aversion in decision-making under risk. *Science* 315:515-518.

27. Hare TA et al. (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J Neurosci* 28:5623-5630.

28. FitzGerald TH, Seymour B, Dolan RJ (2009) The role of human orbitofrontal cortex in value comparison for incommensurable objects. *J Neurosci* 29:8388-8395.

29. Kahnt T, Heinzle J, Park SQ, Haynes JD (2010) The neural code of reward anticipation in human orbitofrontal cortex. *Proc Natl Acad Sci U S A* 107:6010-6015.

30. Plassmann H, O'Doherty JP, Rangel A (2010) Appetitive and aversive goal values are encoded in the medial orbitofrontal cortex at the time of decision making. *J Neurosci* 30:10799-10808.

31. Kahnt T, Heinzle J, Park SQ, Haynes JD (2011) Decoding the Formation of Reward Predictions across Learning. *J Neurosci* 31:14624-14630.

32. Park SQ, Kahnt T, Rieskamp J, Heekeren HR (2011) Neurobiology of value integration: when value impacts valuation. *J Neurosci* 31:9307-9314.

33. De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16:105-110.

34. Klein-Flugge MC et al. (2013) Segregated Encoding of Reward-Identity and Stimulus-Reward Associations in Human Orbitofrontal Cortex. *J Neurosci* 33:3202-3211.

35. McNamee D, Rangel A, O'Doherty JP (2013) Category-dependent and category-independent goal-value codes in human ventromedial prefrontal cortex. *Nat Neurosci*.

36. Lebreton M et al. (2009) An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron* 64:431-439.

37. Leathers ML, Olson CR (2012) In monkeys making value-based decisions, LIP neurons encode cue salience and not action value. *Science* 338:132-135.

38. Wacker P, Deneffe D (1996) Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown. *Management Science* 42(8):1131-1150.

39. Peck CJ et al. (2009) Reward modulates attention independently of action value in posterior parietal cortex. *J Neurosci* 29:11182-11191.

40. Kahnt T, Tobler PN (2013) Salience signals in the right temporoparietal junction facilitate value-based decisions. *J Neurosci* 33:863-869.

41. Gottlieb J (2012) Attention, learning, and the value of information. *Neuron* 76:281-295.

42. Logothetis NK et al. (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412:150-157.

43. Logothetis NK (2008) What we can do and what we cannot do with fMRI. *Nature* 453:869-878.

44. Newsome WT et al. (2013) Comment on "In monkeys making value-based decisions, LIP neurons encode cue salience and not action value". *Science* 340:430.

45. Liu Y, Yttri EA, Snyder LH (2010) Intention and attention: different functional roles for LIPd and LIPv. *Nat Neurosci* 13:495-500.

46. Mitchell, C. J. & Le Pelley, M. E. (2010) *Attention and associative learning: from brain to behaviour* (Oxford University Press, Oxford, UK).

47. Ogawa M et al. (2013) Risk-responsive orbitofrontal neurons track acquired salience. *Neuron* 77:251-258.

48. O'Neill M, Schultz W (2010) Coding of reward risk by orbitofrontal neurons is mostly distinct from coding of reward value. *Neuron* 68:789-800.

49. Haynes JD, Rees G (2005) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci* 8:686-691.

50. Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. *Nat Neurosci* 8:679-685.

51. Haynes JD et al. (2007) Reading hidden intentions in the human brain. *Current Biology* 17:323-328.

52. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci U S A* 103:3863-3868.

53. Kahnt T, Grueschow M, Speck O, Haynes JD (2011) Perceptual learning and decision-making in human medial frontal cortex. *Neuron* 70:549-559.

54. Kahnt T, Heinzle J, Park SQ, Haynes JD (2011) Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. *Neuroimage* 56:709-715.

55. Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI (2009) Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci* 12:535-540.

## Figure Legends

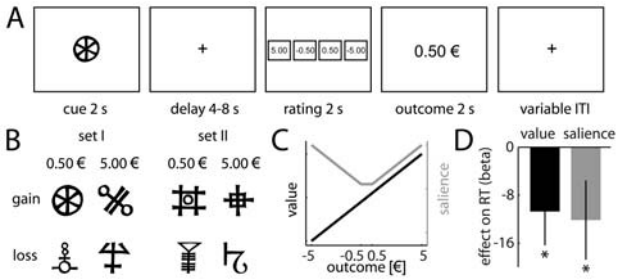


**Fig. 1.** Task structure, stimuli and behavioral results. (*A*) Structure and timing of task. Associations between buttons and ratings were randomized in each trial. (*B*) Example of cue-outcome associations (actual associations were randomized across subjects). (*C*) Dissociation of value and salience using appetitive and aversive outcomes. Salience corresponds to the absolute value of predicted outcomes. Note that the use of only appetitive (or aversive) outcomes alone would not allow to dissociate value and salience. (*D*) Effects of value and salience on response times (RTs) for the ratings. Bars represent standardized regression coefficients from individual multiple regressions, averaged across subjects. Error bars depict SEM for n = 30. Asterisks indicate significant effects at $p < 0.05$ (one sample t-test).

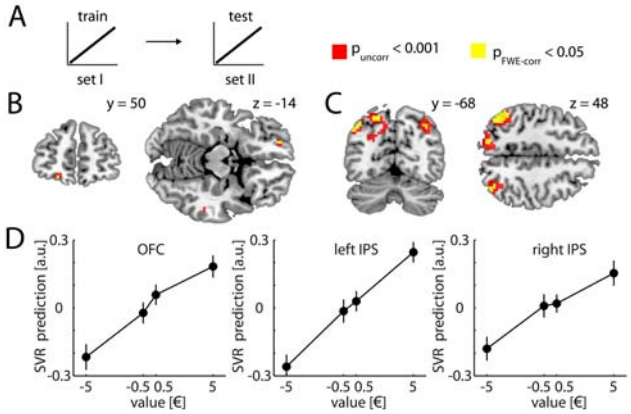


**Fig. 2.** Decoding value information. (*A*) Schematic of the decoding analysis. SVR models were trained on data from set I and tested on data from set II (and vice versa) across appetitive and aversive cues. (*B*). Coronal (*left*) and transversal (*right*) sections depicting regions in the OFC with significant information about the value of the cues. (*C*). Coronal (*left*) and transversal (*right*) sections depicting regions in the superior PPC with significant information about the value of the cues. For display purposes, t-map (one sample t-test) is thresholded at $p < 0.05_{FWE\text{-}corr}$ (yellow), and $p_{uncorr} < 0.001$ (red). (*D*) For illustration proposes, labels predicted by the SVR model are plotted as a function of the actual values in the test data set. SVR outputs from peak searchlights are normalized and averaged across cross-validation steps and subjects. Error bars depict SEM for n = 30.

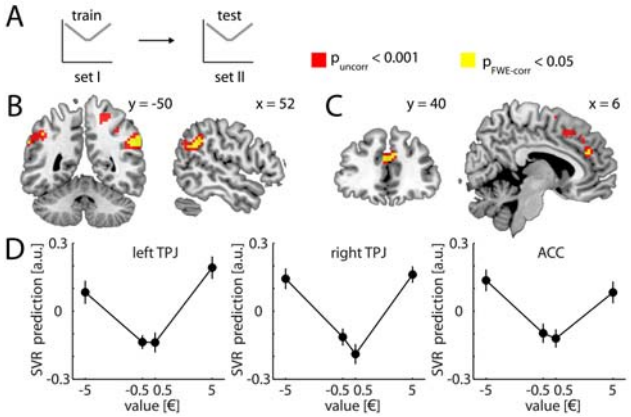


**Fig. 3.** Decoding salience information. (*A*) Schematic of the decoding analysis. SVR models were trained on data from set I and tested on data from set II (and vice versa) across appetitive and aversive cues. (*B*). Coronal (*left*) and sagittal

(*right*) sections depicting regions in the inferior PPC with significant information about the salience of the cues. (*C*). Coronal (*left*) and sagittal (*right*) sections depicting regions in the ACC with significant information about the salience of the cues. For display purposes, t-map (one sample t-test) is thresholded at $p < 0.05_{FWE\text{-}corr}$ (yellow), and $p_{uncorr} < 0.001$ (red). (*D*) For illustration proposes, labels predicted by the SVR model are plotted as a function of the actual values in the test data set. SVR outputs from peak searchlights are normalized and averaged across cross-validation steps and subjects. Error bars depict SEM for n = 30.

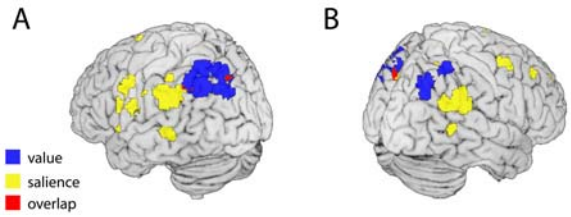


**Fig. 4.** Value and salience signals in the posterior parietal cortex. Surface plots of the left (*A*) and right (*B*) hemisphere depict regions with significant information about value (blue), salience (yellow) and their overlap (red). For illustrative purposes, t-maps (one sample t-tests) are thresholded at $p_{uncorr} < 0.001$.

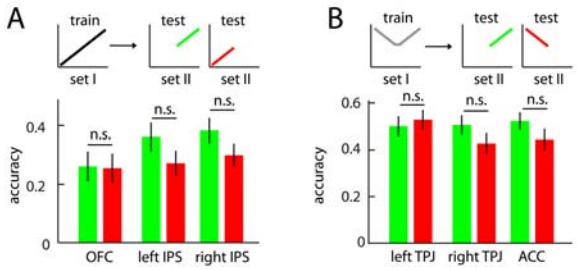


**Fig. 5.** Contribution of appetitive and aversive cues to neural encoding of value and salience. (*A*) Top panel illustrates the decoding analysis. SVR models were trained on the value of all cues from set I and tested on the value of either only appetitive (green) or only aversive cues (red) from set II (and vice versa). Bars reflect average accuracy (Fisher's z-transformed correlation) for appetitive (green bars) and aversive cues (red bars) in the OFC, left and right IPS (superior PPC). All accuracies are significant at $p < 0.001$ (one sample t-test). (*B*) Top panel illustrates the decoding analysis. SVR models were trained on the salience of all

cues from set I and tested on the salience of either only appetitive (green) or only aversive cues (red) from set II (and vice versa). Bars reflect average accuracy for appetitive (green bars) and aversive cues (red bars) in the ACC, left and right TPJ (inferior PPC). All accuracies are significant at $p < 0.001$ (one sample t-test). Error bars depict SEM for n = 30, n.s. indicates non-significant differences (paired t-tests, all ps > 0.19).

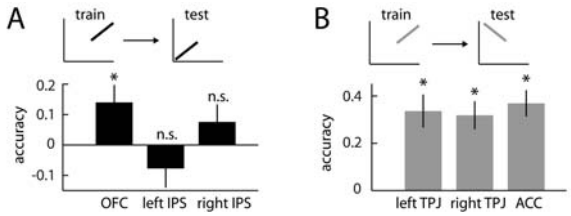


**Fig. 6.** Common value and salience scales for appetitive and aversive cues. (*A*) Top panel illustrates the decoding analysis. SVR models were trained on the value of all appetitive cues and tested on the value of all aversive cues (and vice versa). Bars reflect average accuracy (Fisher's z-transformed correlation) in the OFC, left and right IPS (superior PPC). Accuracies differ significantly between regions (one-way ANOVA; $F_{2,58} = 3.41$, $p = 0.04$). (*B*) Top panel illustrates the decoding analysis. SVR models were trained on the salience of all appetitive cues and tested on the salience of all aversive cues (and vice versa). Bars reflect average accuracy in the ACC, left and right TPJ (inferior PPC). Accuracies do not differ between regions (one-way ANOVA; $F_{2,58} = 0.24$, $p = 0.79$). Error bars depict SEM for n = 30. Asterisks indicate significant accuracy at $p < 0.05$ (one sample t-test) and n.s. indicates non-significance at this threshold (i.e. $p > 0.05$).